Statistics Canada | Statistique Canada

Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

Canada

TRAITEMENT DE LA NON-RÉPONSE DE VAGUE DANS LES ENQUÊTES LONGITUDINALES

Eric Langlet

Working Paper No. SSMD 86-012F

TRAITEMENT DE LA NON-RÉPONSE DE VAGUE

DANS LES ENQUÊTES LONGITUDINALES

Eric Langlet

Août 1986
Division des Méthodes d'Enquêtes Sociales

ABSTRACT

This paper summarizes and compares the different imputation and weighting methods used for wave nonresponse in longitudinal surveys. In addition, several methods used or suggested in various longitudinal surveys conducted in the United States are described.

In most cross-sectional surveys, two methods, weighting and imputation, are used to deal with nonresponse. In general, weighting is used for unit nonresponse, while imputation is used to correct for item nonresponse.

The selection of the method to be used for wave nonresponse in longitudinal surveys is not a simple problem. From a cross-sectional perspective, nonresponse for a particular wave may be viewed as unit nonresponse. However, from a longitudinal perspective, wave nonresponse may be viewed as a set of item nonresponses in the longitudinal record.

TABLE DES MATIÈRES

1. Introduction 1

2. Approche de pondération de Little et David 3

   2.1 Approche de pondération pour schémas emboîtés 3

   2.2 Approche de pondération pour schémas non emboîtés 6

3. Méthodes d'imputation 9

   3.1 Exposé des méthodes d'imputation de Kalton et Lepkovski 10

   3.2 Comparaison des méthodes d'imputation 12

4. Comparaison de la pondération et de l'imputation 15

5. Méthodes de modélisation pour données catégoriques 17

6. Traitement de la non-réponse de vague dans trois enquêtes longitudinales menées aux États-Unis 20

   6.1 Étude comparative de trois méthodes d'ajustement pour la non-réponse de vague dans le NMCES 21

   6.2 Traitement de la non-réponse de vague dans le NMCUES 25

      6.2.1 Hot deck séquentiel pondéré 26

   6.3 Traitement de la non-réponse de vague dans SIPP 27

      6.3.1 Méthode d'imputation pour variables discrètes 27

      6.3.2 Méthodes d'imputation pour variables continues 29

         6.3.2.1 Méthode d'imputation itérative par régression 29

         6.3.2.2 Technique d'appariement 30

      6.3.3 Recherche sur la non-réponse de vague dans SIPP 31

7. Conclusion 32

Bibliographie 33

TRAITEMENT DE LA NON-RÉPONSE DE VAGUE DANS LES ENQUÊTES LONGITUDINALES

1. INTRODUCTION

En général, deux formes d'ajustement pour la non-réponse dans les enquêtes statistiques peuvent être distinguées: d'abord, la pondération qui consiste à assigner aux unités répondantes des poids d'agrandissement de façon à compenser la non-réponse des autres unités; puis, l'imputation, technique qui consiste à assigner certaines valeurs aux données manquantes. Traditionnellement, dans la plupart des enquêtes transversales (en anglais, "cross-sectional surveys"), la pondération est utilisée pour traiter la non-réponse totale alors que l'imputation est choisie pour traiter la non-réponse partielle.

Ce travail s'intéresse à la généralisation de cette stratégie mixte aux enquêtes longitudinales recueillant de l'information sur les mêmes unités statistiques à différents points dans le temps appelés vagues. Dans ce type d'enquête, il est possible de distinguer trois différentes formes de non-réponse: premièrement, la non-réponse totale, lorsqu'un individu (ou toute autre unité statistique telle que la famille, le ménage, etc.) ne répond à aucune des vagues; deuxièmement, la non-réponse de vague (en anglais, "wave nonresponse"), lorsqu'un individu ne répond qu'à certaines vagues; et, troisièmement, la non-réponse partielle, lorsqu'une partie seulement de l'information est manquante lors d'une vague particulière.

Par analogie avec la stratégie adoptée dans les enquêtes transversales, il apparaît logique de traiter dans les enquêtes longitudinales, la non-réponse totale par pondération et la non-réponse partielle par imputation (y compris les méthodes de modélisation). Une question intéressante se pose ici: quelle stratégie devrait-on adopter pour traiter la non-réponse de vague? Devrait-on utiliser la pondération, l'imputation ou une combinaison de ces deux méthodes?

À la section 2, on décrira la méthode de pondération de Little et David (1983), une approche intéressante puisqu'elle est peut-être la seule méthode de pondération valable qui traite la non-réponse de vague dans n'importe quel type d'enquêtes longitudinales. C'est pourquoi on y a consacré une attention particulière.

On présentera à la section 3 plusieurs méthodes d'imputation qui pourraient être utilisées pour traiter la non-réponse de vague. Les méthodes seront ensuite comparées à l'aide d'une expérience de simulation.

La section 4 donne le pour et le contre des approches de pondération et d'imputation dans les enquêtes longitudinales.

On décrira à la section 5 certaines méthodes de modélisation pour données catégoriques utilisant les modèles log-linéaires.

Pour terminer, on verra à la section 6 des méthodes suggérées ou utilisées dans quelques enquêtes longitudinales aux États-Unis.

Ce travail se veut donc une synthèse des différentes méthodes de pondération et d'imputation utilisées pour le traitement de la non-réponse de vague. De plus, dans certains cas, on verra la conséquence de l'application de ces diverses méthodes sur la variance des estimateurs.

2. APPROCHE DE PONDÉRATION DE LITTLE ET DAVID (1983)

Little et David font la distinction entre trois types de non-réponse de vague: la perte, l'entrée répétée et l'entrée tardive. La perte se produit lorsqu'un individu ne répond pas ou n'est pas contacté à une vague particulière et n'est plus interviewé par la suite. Cette forme de non-réponse de vague est la plus courante et résulte en une diminution progressive des unités interviewées formant ainsi un schéma emboîté (en anglais, "nested pattern") de non-réponse de vague. Pour ce type de non-réponse de vague, Little et David suggèrent une méthode relativement simple de pondération: la pondération pour schémas emboîtés (voir sous-section 2.1). L'entrée répétée, quant à elle, a lieu lorsqu'un individu interviewé à certaines vagues, disparaît de l'enquête pour une ou plusieurs vagues, puis réapparaît à une vague subséquente. À noter qu'un individu peut quitter l'enquête plus d'une fois. Finalement, l'entrée tardive représente le cas où un individu non interviewé à la première vague entre à une vague ultérieure pour rester dans l'enquête jusqu'à la fin. Lorsque la non-réponse de vague n'est pas constituée uniquement de perte, Little et David suggère une méthode de pondération appelée pondération pour schémas non emboités (voir sous-section 2.2).

2.1 Approche de pondération pour schémas emboîtés

Dans les enquêtes longitudinales, on fait la distinction entre deux types de poids: les poids transversaux appliqués aux répondants à une vague particulière afin de produire des estimés transversaux et les poids longitudinaux appliqués aux répondants sur une série de vagues pour lesquelles on désire calculer des estimés longitudinaux.

Dans le cas des schémas emboîtés, on peut voir facilement que les poids transversaux à une vague particulière, servent de poids longitudinaux à toute série de vagues se terminant à cette vague, ce qui rend la méthode relativement simple.

L'approche de Little et David tente de tenir compte de toute l'information auxiliaire disponible aux autres vagues. L'objectif de cette méthode est de prédire la réponse ou la non-réponse à la $j^{ième}$ occasion utilisant toute l'information disponible provenant des individus ayant répondu à toutes les vagues précédentes. Les poids sont calculés comme étant inversement

proportionnels à une estimation de la probabilité de réponse, calculée à l'aide de toutes les covariables disponibles.

Afin de simplifier le problème, supposons qu'on n'ait que trois vagues avec n unités échantillonées, dont m_1 répondants à la première occasion, m_2 à la seconde et m_3 à la troisième. Pour chaque unité de l'enquête, on définit un vecteur de réponse (r_1, r_2, r_3) où r_j prend la valeur 1 si l'unité est observée à la vague j et 0 sinon, j = 1, 2, 3. Dénotons par z l'ensemble des variables du plan d'échantillonnage observées pour toutes les n unités. Dénotons par x_j, l'ensemble des variables recueillies pour les m_j répondants à l'occasion j, j = 1, 2, 3.

La méthode de pondération est basée sur les régressions de:

(2.1) r_1 sur z, utilisant toutes les n observations

(2.2) r_2 sur z, x_1, utilisant m_1 observations avec $r_1 = 1$

(2.3) r_3 sur z, x_1, x_2, utilisant m_2 observations avec $r_2 = r_1 = 1$

Les régressions doivent être logistiques ou probits en vertu de la nature dichotomique des indicateurs de réponse. Les unités répondantes à la première vague seront pondérées par w_1 où

$$w_1^{-1} = \hat{P}\left[r_1 = 1 \mid Z\right] ,$$

soit la valeur moyenne estimée de r_1 provenant de la régression (2.1) pour chaque valeur de z. Si les variables du plan d'échantillonnage sont des strates et des unités primaires d'échantillonnage (UPE), ces poids pourraient correspondre à l'inverse du taux de réponse pour chaque combinaison "strate - UPE", par exemple.

Les unités répondantes de la deuxième vague seront pondérées par

$$w_2 = w_1 w_{2.1}$$

où $w_{2.1}^{-1} = \hat{P}\left[r_2 = 1 \mid r_1 = 1, z, x_1\right]$,

la valeur moyenne de r_2 estimée à partir de la régression (2.2) pour chaque valeur de z et x_1. Le poids $w_{2.1}$ représente l'ajustement nécessaire dû à la perte d'unités entre la première et la deuxième vague.

Les unités répondantes de la 3ième vague seront pondérées par

$$w_3 = w_1 w_{2.1} w_{3.12} \quad ,$$

où $w_{3.12}^{-1} = \hat{P}\left[r_3 = 1 \mid r_1 = r_2 = 1,\ z,\ x_1,\ x_2\right]$,

la valeur moyenne de r_3 estimée à l'aide de la régression (2.3) pour chaque combinaison de z, x_1 et x_2. $w_{3.12}$ représente l'ajustement nécessaire dû à la perte d'unités entre la 2ième et la 3ième vague.

Cette approche repose sur l'hypothèse que la réponse à une vague particulière ne dépend pas des données de cette vague et des suivantes une fois qu'on a tenu compte des données aux vagues précédentes. On parlera ici de mécanisme de non-réponse de vague "ignorable" au sens de Little (1982).

Le problème majeur est d'estimer les valeurs moyennes des indicateurs de réponse à chacune des vagues par régression logistique ou probit, tâche qui devient de plus en plus ardue dans les dernière vagues étant donné la quantité ,considérable d'information auxiliaire disponible aux vagues précédentes. Cependant, cette modélisation est beaucoup moins fastidieuse que celle requise pour l'imputation utilisant des modèles multivariés pour prédire tous les items recueillis à chacune des vagues.

2.2 Approche de pondération pour schémas non emboîtés

La méthode utilisée pour les schémas emboîtés est modifiée en ajoutant d'autres régressions de façon à tenir compte de tous les schémas possibles de non-réponse. Afin d'illustrer la méthode, considérons d'abord un schéma non emboîté à deux vagues seulement.

Supposons que les unités soient ordonnées de façon à ce que les m premières soient observées pour les deux vagues ($r_1 = r_2 = 1$), les $m_1 - m$ suivantes soient observées à la première vague seulement ($r_1 = 1$, $r_2 = 0$), les $m_2 - m_1$ suivantes aient répondu à la seconde vague seulement et les dernières $n - m_2$ unités n'aient répondu à aucune des deux vagues. En conservant les mêmes définitions de z, x_1 et x_2 que dans la section précédente, les poids sont maintenant basés sur les régressions de:

(2.4) r_1 sur z, utilisant n observations

(2.5) r_2 sur z, x_1, utilisant m_1 observations avec $r_1 = 1$

(2.6) r_2 sur z, utilisant $n - m_1$ observations avec $r_1 = 0$

Les m_1 unités répondantes de la première vague seront pondérées par w_1, où

$$w_1^{-1} = \hat{P}\left[r_1 = 1 \mid z\right] ,$$

la valeur moyenne estimée de r_1 provenant de la régression (2.4) pour chaque valeur de z.

Les unités 1 à m ayant répondu à la deuxième vague seront pondérées par $w_{2.1}$, où

$$w_{2.1}^{-1} = P\left[r_2 = 1 \mid r_1 = 1,\ z,\ x_1\right] ,$$

la valeur moyenne estimée de r_2 provenant de la régression (2.5). Les m unités pondérées représentent donc les m_1 premières unités. Les unités $m_1 + 1$ à m_2 ayant répondu à la deuxième vague seront pondérées par $w_{2.1}$, où

$$w_{2.1}^{-1} = P\left[r_2 = 1 \mid r_1 = 0,\ z\right],$$

la valeur moyenne estimée de r_2 provenant de la régression (2.6). Ces $m_2 - m_1$ unités pondérées à la deuxième vague représentent donc les $n - m_1$ dernières unités. Les poids transversaux des vagues 1 et 2 seront w_1 et $w_{2.1}$ respectivement, alors que les poids longitudinaux combinant les deux vagues seront obtenus par le produit de w_1 par $w_{2.1}$.

À noter que le choix des régressions était quelque peu arbitraire. L'ordre des régressions aurait pu être renversé complètement de façon à pondérer les unités répondantes de la seconde vague en faisant une régression de r_2 sur z, puis pondérer les unités répondantes de la première vague pour lesquelles $r_2 = 1$ en faisant une régression de r_1 sur z et x_2 et finalement pondérer les unités répondantes de la première vague pour lesquelles $r_2 = 0$ en faisant une régression de r_1 sur z. Les poids longitudinaux seraient alors obtenus en multipliant w_2 par $w_{1.2}$ et seraient en général différent du produit de w_1 par $w_{2.1}$.

Le fait que les poids longitudinaux diffèrent selon l'ordre des régressions rend la méthode peu attrayante. De plus, les observations hors du schéma emboîté (dans ce cas-ci les $m_2 - m_1$ unités ayant répondu uniquement à la seconde vague) ont tendance à recevoir des poids élevés (aucune limite sur ces poids), ce qui a pour conséquence d'augmenter la variance des estimateurs malgré la réduction du biais. Pour une discussion sur le problème d'augmentation de la variance dû à la pondération, voir David, Little, Samuhel et Triest (1983).

Maintenant, si on ajoute une troisième vague, on doit considérer quatre nouvelles régressions, soit

(2.7) r_3 sur z, x_1, x_2 utilisant les observations pour lesquelles $r_1 = r_2 = 1$
(2.8) r_3 sur z, x_1, utilisant les observations pour lesquelles $r_1 = 1$ et $r_2 = 0$
(2.9) r_3 sur z, x_2, utilisant les observatios pour lesquelles $r_1 = 0$ et $r_2 = 1$
(2.10) r_3 sur z, utilisant les observations pour lesquelles $r_1 = r_2 = 0$.

Certaines de ces régressions peuvent être inutiles si le schéma correspondant de non-réponse n'est pas observé. Cependant, l'ajout de chaque vague multiplie potentiellement le nombre de régressions par deux, rendant ainsi la méthode laborieuse pour une enquête longitudinale comportant un grand nombre de vagues ($(2^k - 1)$ régressions possibles pour k vagues).

Une illustration simplifiée de ce procédé appliqué au traitement de la non-réponse de vague pour l'enquête "Survey of Income and Program Participation (SIPP)" est décrite dans l'article de David et Little (1982).

En vertu des raisons mentionnées précédemment, il apparaît souhaitable d'adopter une stratégie limitant le nombre d'observations hors du schéma emboîté. David et Little suggèrent comme compromis d'utiliser l'imputation sur les unités pour lesquelles une seule vague manque et d'éliminer, pour les unités restantes hors du schéma emboîté, l'information obtenue à certaines vagues de façon à obtenir un schéma emboîté. Selon eux, la perte d'information devrait être faible, et la méthode de pondération pour schémas emboîtés pourrait alors être utilisée.

En guise de conclusion, les auteurs mentionnent que ces méthodes devraient être testées sur des données réelles afin d'améliorer la méthode si nécessaire et de démontrer l'utilité de cette approche de pondération.

3. MÉTHODES D'IMPUTATION

Une méthode d'imputation pour un item manquant à une vague particulière est dite longitudinale (en anglais, "cross wave imputation") si elle utilise l'information obtenue aux autres vagues. Sinon, la méthode d'imputation est dite transversale (en anglais, "cross-sectional imputation").

Plus les variables auxiliaires des vagues observées améliorent l'estimation d'un item manquant à une vague particulière, plus le gain de l'imputation longitudinale sera grand par rapport à l'imputation transversale. Les variables répétées d'une vague à l'autre se prêtent très bien à l'imputation longitudinale à condition que cette même variable soit disponible à l'une ou l'autre des autres vagues, ce qui n'est pas toujours le cas. Il est donc indispensable de créer une méthode d'imputation transversale comme solution de rechange.

Il existe un problème d'ordre pratique avec l'imputation longitudinale: elle nécessite la création d'un fichier longitudinal, tâche qui peut devenir complexe, en raison des problèmes d'appariement qui peuvent survenir. Ce problème a été particulièrement évident lors de l'appariement des deux premières vagues du programme "Income Survey Development Program" (ISDP) (Lepkowski (1982)). Également, les contrôles doivent être faits longitudinalement de façon à éviter les incohérences entre les vagues. Il y a aussi une question de production: en effet, le meilleur système d'imputation devrait utiliser l'information provenant de toutes les vagues. Ce système ne peut donc être utilisé qu'une fois l'enquête longitudinale terminée, ce qui peut s'avérer très long dans les enquêtes s'étendant sur plusieurs années. Pour fin de production, il faudrait effectuer l'imputation après chaque vague. Ainsi, une fois une nouvelle vague entrée, on devrait théoriquement reprendre les imputations effectuées sur les vagues précédentes puisque cette nouvelle vague ajoute de l'information possiblement utile à l'imputation. L'opération pourrait devenir très coûteuse puisque, de cette façon, une enquête comportant n vagues nécessiterait l'imputation de $n(n + 1)/2$ vagues.

Kalton et Lepkowski (1983) proposent plusieurs méthodes d'imputation longitudinales. Ce sont ces méthodes qu'on décrira à la section suivante pour ensuite les comparer à la section 3.2.

3.1 Exposé des méthodes d'imputation de Kalton et Lepkowski

Dans la discussion qui va suivre, supposons qu'on n'ait que deux vagues. Pour le $i^{ième}$ individu, on désire imputer une valeur à la variable y_{2i}, la valeur manquante de la variable y à la deuxième occasion et supposons que y_{1i}, la valeur de la variable y à la première occasion soit observée.

(a) Substitution directe

On substitue à l'item manquant de la $2^{ième}$ occasion, le même item à la première occasion, c'est-à-dire $\hat{y}_{2i} = y_{1i}$. Cette méthode est valable pour les items très stables d'une vague à l'autre, mais sous-estime évidemment les mesures de changement entre les deux vagues.

(b) Imputation longitudinale hot deck

Il s'agit d'une généralisation de la méthode dite du "hot deck" telle que traditionnellement utilisée où l'on ajoute simplement aux variables auxiliaires définissant les classes d'imputation, la variable d'intérêt observée à la première occasion. Ainsi, si on oublie pour l'instant les autres variables auxiliaires, on obtient $\hat{y}_{2i} = y_{2k}$ où les valeurs y_{1i} et y_{1k} correspondent ou sont près l'une de l'autre et où le $k^{ième}$ individu est utilisé comme donneur. Cette technique fonctionne bien lorsque la variable y est discrète avec peu de catégories. Si cette variable est continue, il faudra regrouper ses valeurs en classes, ce qui résultera en une perte d'information. Dans cette circonstance l'une ou l'autre des méthodes suivantes serait peut-être préférable.

(c) Imputation longitudinale de changement hot deck

On observe chez le donneur la quantité $y_{2k} - y_{1k}$ et on estime y_{2i} par $\hat{y}_{2i} = y_{1i} + (y_{2k} - y_{1k})$. Ce changement est additif. Avec ce modèle, on suppose que la grandeur du changement ne dépend pas de la valeur y_{1i}. Pour combler cette lacune, on peut considérer un modèle de changement

proportionnel où l'on calcule chez le donneur la quantité y_{2k}/y_{1k} et l'on estime y_{2i} par $\hat{y}_{2i} = y_{1i}y_{2k}/y_{1k}$.

A noter que les méthodes (b) et (c) sont stochastiques, dû au fait que, pour chaque imputation, un donneur est sélectionné aléatoirement à l'intérieur de la classe d'imputation. Toutes ces méthodes peuvent devenir déterministes. Pour (b), il suffit d'imputer la moyenne dans la classe d'imputation tandis qu'en (c), on peut calculer le changement moyen (additif ou proportionnel) à l'intérieur de la classe d'imputation. Ainsi, en (b), tous les individus à l'intérieur d'une classe se verront assigner la même valeur tandis qu'en (c), tous les individus se trouvant dans une même classe d'imputation, se verront assigner le même changement.

(d) Imputation longitudinale par la régression

On estime y_{2i} à l'aide de l'équation de régression suivante: $\hat{y}_{2i} = a + by_{1i} + e_i$, où a et b sont les coefficients de régression calculés à l'aide des répondants aux deux vagues et où e_i désigne un résidu quelconque associé à $\hat{y}_{2i}$. Cette méthode est stochastique mais devient déterministe si $e_i = 0$. Les résidus peuvent être déterminés de plusieurs façons. Notons que si b = 1, la méthode correspond au modèle additif de l'approche (c) alors que si a = 0, la méthode correspond au modèle proportionnel de l'approche (c).

Après avoir exposé ces diverses méthodes, il serait utile d'émettre quelques commentaires. En premier lieu, toutes ces méthodes peuvent être raffinées en appliquant le même procédé à l'intérieur du croisement de plusieurs variables catégoriques, si l'on croit, par exemple, que le changement additif ou proportionnel de l'approche (c) sera différent d'une classe à l'autre. En second lieu, ces méthodes peuvent être généralisées à plus de deux vagues en utilisant des modèles de régression plus complexes. En troisième lieu, l'imputation longitudinale peut s'effectuer dans les deux directions. On peut également prédire les réponses à la première vague à partir des réponses à la deuxième vague. Finalement, soulignons encore une fois que ces méthodes ne sont valables que si la variable nécessitant l'imputation est observée à une autre vague. Dans le cas contraire, on

devrait utiliser d'abord une approche transversale pour imputer une valeur à une première vague et ensuite utiliser une approche longitudinale pour traiter la non-réponse aux autres vagues de façon à préserver la structure longitudinale des données.

3.2 Comparaison des méthodes d'imputation

Kalton et Lepkowski ont comparé sept différentes méthodes d'imputation par simulation à l'aide des données des deux premières vagues de l'enquête longitudinale "Income Survey Development Program" (ISDP), enquête préparatoire à l'enquête longitudinale "Survey of Income and Program Participation" (SIPP) dont il sera question à la section 5. Les méthodes comparées sont

(a) la substitution directe
(b) l'imputation longitudinale "hot deck"
(c) l'imputation longitudinale de changement additif "hot deck"
(d) l'imputation longitudinale de changement proportionnel "hot deck"
(e) l'imputation déterministe de changement proportionnel
(f) l'imputation déterministe par la régression
(g) l'imputation transversale "hot deck"

La seule variable qu'ils ont étudiée est le tarif horaire de paie, variable ayant une corrélation de 0.97 entre les deux vagues. Les résultats de leur étude sont donc fort limités.

À partir des deux premières vagues de l'enquête ISDP, Kalton et Lepkowski ont simulé des données manquantes sur les données complètes de façon à refléter le plus possible les schémas de non-réponse observés dans l'échantillon complet. L'expérience de simulation a été répétée 10 fois de façon à rendre les résultats plus stables. On ne fera mention ainsi que des principaux résultats.

L'équation de régression du tarif horaire de paie à la vague 1 (y_1) sur le tarif horaire de paie à la vague 2 (y_2) est donnée par

$$y_1 = 0.20 + 0.985y_2$$

en utilisant les unités ayant répondu aux deux vagues. Comme l'ordonnée à l'origine de cette équation est près de 0, on s'attend à ce que les méthodes (e) et (f) produisent des résultats similaires. De plus, comme la pente de la droite de régression est près de l'unité, ces deux dernières méthodes devraient fournir des résultats voisins de la substitution directe. Ces deux faits combinés impliquent également que les méthodes d'imputation (c) et (d) devraient fournir des résultats semblables.

Deux mesures de qualité de l'ajustement de l'imputation ont été calculées, soient une mesure du biais ainsi qu'une mesure de précision. Ces mesures ont été prises à chacune des deux vagues. Aucune des méthodes n'a produit de mesure du biais significativement différente de 0. Comme on pouvait s'y attendre, toutes les méthodes d'imputation longitudinales ont fourni de meilleurs résultats que l'approche transversale en termes de précision de l'imputation. Notons, à cet effet, que lorsqu'on compare l'imputation longitudinale et transversale, on compare en fait une combinaison d'imputations longitudinales et transversales à l'imputation transversale, puisque, dans bien des cas, l'approche longitudinale ne peut être utilisée. Il en résulte que l'amélioration causée par l'emploi de l'imputation longitudinale se trouve sous-estimée. Toujours en termes de précision de l'imputation, toutes les approches déterministes ont produit de meilleurs résultats que les approches stochastiques. Cependant, les approches déterministes, contrairement aux approches stochastiques, transforment les lois de probabilité sous-jacentes aux données.

Une autre étude de qualité de l'ajustement a été effectuée sur la différence du tarif horaire de paie entre les deux vagues. L'avantage des méthodes longitudinales a été encore plus marqué dans cette étude. L'écart-type de cette différence a été calculé pour chacune des méthodes pour ensuite être comparé à la valeur réelle. Toutes les méthodes surestiment cette valeur réelle, en particulier l'approche transversale qui double la vraie valeur. Encore une fois, l'utilisation de la méthode transversale lorsque les réponses aux deux vagues étaient manquantes, expliquent pourquoi les approches longitudinales ont également surestimé cet écart-type.

Un des problèmes majeurs de ces approches longitudinales, est que celles-ci traitent les variables une à la fois. Par contre, la méthode du hot deck transversale, par exemple, peut traiter tout un groupe de variables en même temps en choisissant un seul donneur, préservant ainsi les relations à l'intérieur de ce groupe de variables. Ainsi, l'utilisation de ces méthodes d'imputation pour traiter la non-réponse de vague semble difficile et le problème de trouver une méthode d'imputation valable pour traiter la non-réponse de vague demeure un problème ouvert.

4. COMPARAISON DE LA PONDÉRATION ET DE L'IMPUTATION

Il y a deux façons de considérer la non-réponse de vague dans les enquêtes longitudinales. Vue de façon transversale, la non-réponse de vague n'est rien d'autre que de la non-réponse totale, et la pondération semble appropriée pour traiter ce type de non-réponse. Par contre, vue de façon longitudinale, la non-réponse de vague peut être considérée comme un ensemble d'items manquants qui peuvent être traités par imputation. Ainsi, les méthodes d'imputation pour traiter la non-réponse partielle dans les enquêtes longitudinales peuvent être également utilisées ou adaptées pour la non-réponse de vague.

L'avantage de la pondération pour traiter la non-réponse de vague est qu'elle permet de traiter toutes les variables d'un seul coup pour une unité statistique n'ayant pas fourni de réponse à une occasion particulière alors que l'imputation nécessite l'estimation d'une quantité considérable d'information dans la plupart des enquêtes longitudinales.

Cependant, selon Little et David (1983), l'imputation basée sur un modèle reliant les items manquants d'une vague particulière à l'information obtenue à partir des vagues observées serait théoriquement supérieure à la pondération, même pour un schéma parfaitement emboîté. Les items répétés et très corrélés d'une vague à l'autre se prêtent très bien à l'imputation. En effet, pour un individu quelconque n'ayant pas répondu à une vague particulière, ces items peuvent être prédits de façon précise à l'aide des vagues observées de ce même individu. Ainsi, plus on a de vagues non manquantes, plus on a d'information auxiliaire et plus l'avantage de l'imputation sur la pondération devrait être grand.

Le problème majeur de l'imputation est que, contrairement à la pondération, la relation entre les variables d'une même vague n'est souvent pas préservée. L'imputation va généralement préserver la structure de covariance entre la variable imputée et les variables auxiliaires utilisées lors de l'imputation alors que la covariance avec les autres variables se trouvera atténuée. Bien souvent, à moins que des

précautions soient prises (par exemple, imputation de groupes de variables en même temps, éditer de nouveau les données imputées, etc.), l'imputation peut mener à des incohérences entre les variables ce qui constitue un sérieux désavantage lorsqu'on désire analyser la relation entre les variables à l'aide de régressions ou de tableaux de contingence. De plus, l'imputation fabrique des données, et bien des utilisateurs se servent du fichier imputé et mesurent l'erreur d'échantillonnage en considérant les données imputées comme réelles, surestimant ainsi la précision des estimateurs. Toutefois, ce problème est diminué considérablement si les valeurs imputées sont près des valeurs réelles, ce qui sera le cas si l'item manquant à une vague a été observé à une ou plusieurs autres vagues, et que ce même item est très corrélé d'une vague à l'autre.

Kalton (1985) présente une excellente discussion de ce dilemme.

5. MÉTHODES DE MODÉLISATION POUR DONNÉES CATÉGORIQUES

La modélisation consiste à formuler un modèle pour les données incomplètes et le mécanisme de non-réponse. Les paramètres de ce modèle sont ensuite estimés par la méthode du maximum de vraisemblance ou toute autre méthode d'estimation. Les méthodes d'imputation et de pondération permettent d'obtenir un ensemble de données sous forme rectangulaire. Cet ensemble de données ainsi transformé peut être ensuite analysé par des méthodes pour données complètes. Par contre, la modélisation analyse les données sous forme incomplète.

Il serait bon de mentionner que les différents modèles étudiés dans cette section ne traitent qu'une variable à la fois. L'extension de ces résultats au traitement de la non-réponse de vague demeure un problème ouvert.

Little (1985) a développé des modèles log-linéaires hiérarchiques tenant compte du mécanisme de non-réponse dans le cas d'une variable catégorique mesurée sur les mêmes individus à deux points dans le temps. Cette variable peut être manquante à l'une ou l'autre des occasions. Le cas où la variable est manquante aux deux occasions peut être traité par pondération.

Supposons que la variable catégorique A peut prendre J valeurs numérotées de 1 à J. Dénotons par A_i, la valeur de cette variable à la $i^{ième}$ occasion et par R_i, l'indicateur de réponse à la $i^{ième}$ occasion, i = 1, 2. On obtient le tableau de contingence à deux facteurs avec deux marginales supplémentaires suivant:

A_2

A_1	1	2		J
1	M_{11}	M_{12}	...	M_{1J}
2	M_{21}	M_{22}	...	M_{2J}
⋮	⋮	⋮		⋮
J	M_{J1}	M_{J2}	...	M_{JJ}

$(R_1 = R_2 = 1)$

N_{11}
N_{12}
⋮
N_{1J}

$(R_1 = 1, R_2 = 0)$

	N_{21}	N_{22}	...	N_{2J}

$(R_1 = 0, R_2 = 1)$

où M_{jk} représente le nombre de répondants aux deux vagues pour lesquels $A_1 = j$ et $A_2 = k$, j, k = 1, ..., J et $N_{i\ell}$ représente le nombre de répondants à la $i^{ième}$ vague n'ayant pas répondu à l'autre vague pour lesquels $A_i = \ell$, i = 1, 2 et ℓ = 1, ..., J.

On suppose que les comptes dans chaque cellule M_{ij} admettent une loi de probabilité multinômiale avec probabilités $\{\Pi_{jk} = j, k = 1, ..., J\}$. Ces probabilités peuvent être estimées par méthode de maximum de vraissemblance en utilisant des algorithmes relativement simples permettant de distribuer les marginales supplémentaires dans le tableau de contingence. Little a développé plusieurs modèles log-linéaires hiérarchiques à quatre facteurs (A_1, A_2, R_1 et R_2) pour estimer ces probabilités.

En principe, on pourrait généraliser les méthodes de Little à plus de deux occasions. Pour trois occasions, par exemple, on obtiendrait un tableau de contingence à trois dimensions avec trois tableaux marginaux à deux

dimensions et trois tableaux marginaux unidimensionnels. Les modèles log-linéaires comporteraient six facteurs. En général, pour n occasions on obtient un tableau de contingence à n dimensions avec $\binom{n}{i}$ tableaux marginaux à n-i dimensions, i = 1, ..., n-1 soit $\sum_{i=1}^{n} \binom{n}{i}$ tableaux en tout. Les modèles log-linéaires comporteraient 2n facteurs. Les modèles deviennent donc rapidement très complexes.

Little spécifie que ces méthodes pourraient être appliquées à des sous-groupes d'unités définis par le croisement de plusieurs variables démographiques mesurées à la première vague en incorporant ces variables comme nouvelles dimensions dans la table. Cependant, ce procédé serait très complexe et laborieux. Dans ce contexte, il serait beaucoup plus simple d'utiliser une méthode de post-stratification à l'aide de ces variables démographiques.

Binder (1986) a développé une classe de modèles log-linéaires très riches qui peuvent être utilisés pour traiter la non-réponse dans un tableau de contingence multidimensionnel. Ces méthodes peuvent donc être utilisées pour le traitement de la non-réponse dans une enquête longitudinale comportant n vagues en utilisant les différentes vagues comme dimensions du tableau de contingence.

Fay (1986) a également développé des modèles log-linéaires permettant d'expliquer les mécanismes de non-réponse sur des variables catégoriques. Ces modèles, de même que ceux développés par Little constituent en fait des cas particuliers de ceux développés par Binder.

6. TRAITEMENT DE LA NON-RÉPONSE DE VAGUE DANS TROIS ENQUÊTES LONGITUDINALES MENÉES AUX ÉTATS-UNIS

Dans cette section, on présentera les méthodes utilisées ou suggérées pour traiter la non-réponse de vague dans deux enquêtes sur les soins de santé: l'enquête "National Medical Care and Expenditure Survey" (NMCES) et l'enquête "National Medical Care Utilization and Expenditure Survey" (NMCUES), ainsi qu'une enquête économique: l'enquête "Survey of Income and Program Participation" (SIPP).

Une discussion complète des différents aspects du NMCES et du NMCUES se trouve dans Cox et Cohen (1985).

Le NMCES est une enquête ayant pour but de recueillir de l'information sur la santé des résidents des États-Unis, de voir, entre autres, où et comment les soins de santé ont été reçus, de voir quel a été le coût de ces services et comment ils ont été payés. La partie longitudinale de cette enquête porte sur la sélection de 14 000 ménages interviewés à 6 reprises à intervalles d'environ 12 semaines commençant au début de 1977 pour se terminer au milieu de 1978.

Le NMCUES, successeur du NMCES, est une enquête dont le but, similaire au NMCES, est de fournir des données sur l'utilisation et les dépenses en matière de soins de santé chez les résidents des États-Unis en milieu non hospitalier avec un intérêt particulier pour les personnes couvertes par les régimes d'assurance-maladie "Medicare" et "Medicaid". Cette enquête comporte deux composantes longitudinales. La première, l'enquête "National Household Survey" (NHS), est constituée d'un échantillon stratifié en grappes de 7 200 logements, résultant en 6 600 ménages répondant à la première vague. Ces ménages ont été interviewés à 5 occasions à approximativement 12 semaines d'intervalle commençant au début de 1980 pour se terminer au milieu de 1981. La seconde composante, l'enquête "State Medicaid Household Survey" (SMHS), est constituée d'un échantillon de 4 000 ménages couverts par Medicaid répartis également entre quatre états. Cette enquête comporte également 5 occasions à 12 semaines d'intervalle.

SIPP est une enquête longitudinale portant sur la situation économique des ménages vivant aux États-Unis. Chacun des 26 000 ménages contenus dans l'échantillon a été interviewé à 8 reprises à environ 4 mois d'intervalle, pour une période d'environ 2 ans et demi commençant en octobre 1983.

Dans la section suivante, trois approches de compensation pour la perte d'unités dans le NMCES seront analysées (Cohen (1982)). Dans la section 6.2 on verra la méthode utilisée pour corriger la non-réponse de vague dans le NMCUES. À la section 6.3, on verra des méthodes suggérées pour le traitement de la non-réponse de vague dans SIPP.

6.1 Étude comparative de trois méthodes d'ajustement pour la non-réponse de vague dans le NMCES

En plus d'un taux de non-réponse totale de 9%, 11% des participants au NMCES n'ont répondu qu'en partie pour la période de temps durant laquelle ils étaient éligibles. On désignera ces individus par répondants partiels. Dans l'optique de ce travail, on peut considérer ce type de non-réponse comme non-réponse de vague. Une étude comparative du profil démographique des répondants complets et des répondants partiels a révélé des différences significatives entre ces deux groupes. Trois méthodes d'ajustement ont été suggérées afin de réduire le biais introduit par ce type de non-réponse.

Ces méthodes ont été comparées à l'aide d'une expérience de simulation où l'on a transformé une partie des répondants complets en répondants partiels. Dans cette expérience, les 4 146 répondants partiels ont été appariés aux 34 669 répondants complets. Comme les répondants partiels avaient en général un profil démographique différent des répondants complets, les variables suivantes ont été choisies pour l'appariement: état de santé, âge, race, sexe, région de recensement, assurance ainsi que le nombre de médicaments prescrits durant la période pour laquelle les répondants partiels ont répondu. On obtient ainsi un fichier artificiel contenant 30 523 répondants complets et 4 146 répondants partiels pour lesquels les données sont connues durant leur période de "non-réponse".

Les variables comparées lors de cette expérience portent sur l'utilisation et les dépenses reliées aux médicaments prescrits.

La première méthode envisagée est une méthode de pondération des données des répondants partiels par classe d'ajustement. Cette méthode consiste à modifier les poids des répondants partiels par un facteur d'agrandissement pour chacune des classes d'ajustement définies par le croisement de variables démographiques.

Dénotons par d_i le nombre de jours pour lesquels le $i^{ième}$ participant est éligible et par d_i^* le nombre de jours pour lesquels celui-ci a répondu alors qu'il était éligible. Dans la $g^{ième}$ classe, les poids échantillonaux des répondants partiels ajustés pour la non-réponse totale seront multipliés par le facteur d'ajustement

$$A(g) = \Sigma_{i \in g} w_i d_i \; / \; \Sigma_{i \in g} w_i d_i^*$$

où la sommation porte sur tous les répondants partiels de cette classe et où w_i désigne le poids échantillonnal ajusté pour la non-réponse totale du $i^{ième}$ répondant partiel. L'hypothèse implicite émise ici est qu'à l'intérieur de chacune des classes d'ajustement, le comportement des répondants partiels est uniforme pendant leurs périodes de réponse et de non-réponse. On désignera par la suite cette méthode sous le nom de méthode de pondération des répondants partiels.

La méthode suivante est une méthode d'imputation par appariement des répondants partiels aux répondants complets. Pour l'expérience de simulation, les 4 146 répondants partiels artificiels ont été appariés aux 30 523 répondants complets restants utilisant les mêmes variables ayant servi lors du premier appariement. Les données pour les périodes manquantes des répondants partiels sont alors imputés par les données correspondantes des répondants complets selectionnés lors de l'appariement. La méthode correspond donc à une forme de hot deck longitudinal.

La troisième méthode analysée pourrait être valable si le taux de non-réponse de vague est très faible. Il s'agit d'une méthode de pondération des répondants complets par classe d'ajustement. Les classes choisies sont les mêmes que pour la méthode de pondération des répondants partiels. L'information fournie par les répondants partiels est donc éliminée. On émet donc ici l'hypothèse que les non-répondants partiels et totaux se comportent de la même façon que les répondants complets à l'intérieur de chacune des classes d'ajustement. Dans l'expérience de simulation, les poids échantillonnaux corrigés pour la non-réponse totale, w_i, sont multipliés par le facteur d'agrandissement

$$R(g) = \Sigma_{i \in g,T} w_i \; / \; \Sigma_{i \in g,C} w_i$$

dans la $g^{ième}$ classe d'ajustement où la sommation au numérateur porte sur tous les répondants initiaux (complets et partiels), alors que la sommation au dénominateur porte sur les répondants complets seulement.

Dans l'expérience de simulation, les estimés de différents paramètres utilisant les non-répondants artificiels ont été comparés aux mêmes estimés basés sur les 34 669 répondants complets du fichier original. Ces comparaisons ont été effectuées pour chacune des trois méthodes d'ajustement pour la non-réponse. Les variances de tous les estimateurs ont été évaluées à l'aide de la méthode d'estimation de variance par linéarisation en série de Taylor.

Les résultats de cette expérience peuvent se résumer comme suit: pour les trois méthodes, les estimés des mesures d'utilisation et de dépenses exprimés en termes de moyennes sur des domaines n'ont présenté que très peu de différences significatives lorsque comparés aux estimés calculés à l'aide de tous les répondants complets au niveau $\alpha = 0,05$ (tests t de Student). La seule différence notable a été observée pour l'estimation de la distribution des individus selon leur utilisation des médicaments prescrits, lorsque la méthode de pondération des répondants partiels a été utilisée.

Même si la méthode de pondération des répondants complets a fourni des estimés raisonnablement non-biaisés pour la plupart des paramètres, la destruction d'un 11% additionnel de l'échantillon présente plusieurs inconvénients. En effet, les profils démographiques des répondants partiels et complets diffèrent significativement, ce qui contredit l'hypothèse requise à l'utilisation de cette méthode. Pour les paramètres étudiés dans cette expérience, il semblerait que les classes d'ajustement aient assez bien corrigé la différence entre répondants partiels et complets. Cependant, d'autres paramètres reliés aux mesures d'utilisation et de dépenses en matière de soins de santé pourraient être plus sensibles à cette différence. Finalement, l'élimination des répondants partiels limite la possibilité d'étudier ce sous-groupe dans une analyse subséquente.

L'imputation s'est avérée la meilleure en termes de précision des estimés. De plus, à cause de la procédure d'appariement, l'hypothèse d'homogénéité entre répondants partiels et leurs répondants complets respectifs semble très raisonnable. Cependant, lorsque les questions de coûts et de temps entrent en jeu, cette méthode perd ses avantages. L'appariement lui-même est assez simple, mais la substitution de l'information fournie par les répondants complets sur les périodes de non-réponse des répondants partiels pour chacune des variables liées au temps, représente une tâche très laborieuse.

Comme pour cette enquête, la proportion de répondants partiels était faible, une méthode d'estimation était avantageuse si elle était simple à dériver, était directement applicable à diverses mesures d'analyse reliées au temps, incluait les répondants partiels dans l'analyse sans diminuer la qualité des estimés, et finalement, permettait la construction d'estimés statistiquement valables de mesures d'utilisation et de dépenses exprimés en termes de moyennes et de totaux. Pour ces critères, la méthode de pondération des répondants partiels a été choisie comme étant la meilleure approche pour corriger la non-réponse de vague. Cependant, pour les raisons mentionnées précédemment, la méthode de pondération des répondants complets devrait être utilisée afin d'estimer la distribution

des individus selon leur utilisation des médicaments prescrits. (On rejette ici la méthode d'imputation, parce que trop laborieuse.)

En conclusion, Cohen suggère que d'autres mesures reliées aux soins de santé devraient être étudiées afin de déterminer la meilleure approche pour cette enquête.

6.2 Traitement de la non-réponse de vague dans le NMCUES

Pour une discussion plus complète de ce qui va suivre, le lecteur est référé à Cox (1983) ou à Cox et Sweetland (1982).

Dans cette enquête, les taux de non-réponse à la première vague ont été d'environ 9% pour le NHS et de 3% à 20% pour les quatre SMHS. Ces individus ne devaient plus être interviewés par la suite et une méthode d'ajustement par pondération a été utilisée pour traiter cette non-réponse totale.

Pour l'ensemble des 5 vagues, environ 5% des répondants initiaux ont été perdus pour le NHS comparativement à 4% à 9% pour le SMHS. Cette perte d'individus ne peut être ignorée puisque les analyses du NMCUES utilisent les données cumulées sur toute l'année. Deux méthodes d'ajustement ont été envisagées, soit la pondération et l'imputation. La pondération des unités ayant répondu pour toute la période de l'enquête nécessite l'élimination de l'information des répondants partiels. D'autre part, l'imputation sur les périodes de non-réponse des répondants partiels risque d'introduire beaucoup de variabilité si certaines unités n'ont pas répondu pour la majorité de l'année.

La solution du NMCUES à ce dilemme a été de combiner les deux approches. On rejette l'information obtenue sur les unités ayant répondu pour moins du tiers de l'année et ont pondère le reste des unités. Les autres répondants partiels ont été traités par imputation.

Il s'agit de remplacer les périodes manquantes des répondants partiels par les données obtenues de répondants complets corespondant à ces périodes. La méthode d'imputation choisie est celle du hot deck séquentiel pondéré (En anglais, "weighted sequential hot deck", Cox (1980)), méthode qui fera l'objet de la section suivante.

6.2.1 Hot deck séquentiel pondéré

Cette technique d'imputation est une variante du hot-deck traditionnel. L'algorithme de cette méthode est conçu de façon à ce que les moyennes et proportions estimées à l'aide des données complétées par imputation soient les mêmes en espérance que les moyennes et proportions estimées par pondération utilisant les répondants complets seulement. Cette propriété est également maintenue pour les moments d'ordre supérieur. Comme le hot deck traditionnel, l'imputation s'effectue à l'intérieur de classes d'imputation pour lesquelles les données sont soit aléatoirement manquantes, ou bien, les données manquantes peuvent être expliquées à l'aide d'une fonction de distance associée à une ou plusieurs variables quantitatives (en anglais, "within class sorting variables"). Ainsi, si le fichier est trié par cette ou ces variables quantitatives à l'intérieur de chaque classe, les enregistrements adjacents dans le fichier auront des réponses similaires.

Un des défauts du hot deck séquentiel traditionnel est que cette méthode ne tient pas compte des poids d'analyse des donneurs et des receveurs. De plus, cette méthode ne contrôle pas le nombre de fois qu'une unité est utilisée comme donneur. En effet, un même répondant peut être utilisé plusieurs fois comme donneur puisque les données de ce répondant seront imputées à tous les enregistrements des non-répondants subséquents jusqu'à ce qu'on ait rencontré un autre répondant dans le fichier. La méthode du hot deck séquentiel pondéré, quant à elle, tient compte du nombre de fois qu'un répondant est choisi comme donneur. A l'intérieur de chacune des classes, le nombre de fois, n(i), que les données du $i^{ième}$ répondant seront utilisées pour imputer est une fonction du poids échantillonnal de ce répondant et des poids échantillonnaux des non-répondants pour lesquels ce répondant peut servir de donneur. Ainsi, chaque répondant a une chance d'être choisi comme donneur.

Un des désavantages du hot deck séquentiel pondéré est que cette méthode nécessite quatre passes du fichier de données comparativement à deux pour le hot deck traditionnel.

6.3 Traitement de la non-réponse de vague dans SIPP

Huggins, Samuhel et Weidman (1985) ont développé des méthodes d'imputation applicables à SIPP de même qu'à n'importe quelle enquête longitudinale. Ces méthodes ont été conçues à partir des données du programme "Income Survey Development Program" (ISDP), enquête longitudinale devant préparer SIPP. Dans cette enquête, chaque ménage est interviewé tous les trois mois à six reprises. A chaque interview, les données sont recueillies pour chacun des trois mois précédents. Huggins Samuhel et Weidman ont limité leur recherche à l'imputation d'une seule variable pour une période d'un an (douze occasions). Cependant, ces auteurs mentionnent que de nouvelles recherches sont en cours qui tiendraient compte des autres variables de l'enquête de façon à maintenir la cohérence transversale. Toutes les procédures décrites dans cette section peuvent être raffinées en appliquant chacune d'elles à l'intérieur de classes d'imputation.

6.3.1 Méthode d'imputation pour variables discrètes

Cette méthode est fort simple. Elle consiste à apparier chaque enregistrement ayant un ou plusieurs mois manquants (enregistrement partiel) avec tous les enregistrements complets correspondant avec l'enregistrement partiel pour les mois non-manquants de ce dernier. À l'aide de ces enregistrements complets, on établit la loi de probabilité conditionnelle empirique des valeurs obtenues de la variable d'intérêt dans les mois manquants étant donné les réponses obtenues aux mois non-manquants. Utilisant cette distribution, on choisit aléatoirement un donneur qui servira à compléter les mois manquants de l'enregistrement

partiel. À noter qu'on peut modifier cette méthode en n'exigeant pas un appariement exact. On peut choisir un donneur "suffisamment près" à l'aide d'un critère quelconque. Cette méthode correspond en fait à une forme de hot-deck longitudinal.

Dénotons par $\underline{X} = (X_1, ..., X_{12})$, le vecteur aléatoire des réponses possibles d'un enregistrement longitudinal et dénotons par $\underline{x} = (x_1, ..., x_{12})$ un enregistrement longitudinal quelconque où x_i est une valeur prise par la variable X_i, $i = 1, ..., 12$. De plus, définissons $N(\underline{X} = \underline{x})$ comme étant le nombre de fois que $\underline{x}$ apparaît dans les données complétées par imputation et n comme étant le nombre total d'enregistrements. Alors, en utilisant la méthode mentionnée ci-dessus, on peut montrer que $N(\underline{X} = \underline{x})/n$ dans le fichier des données imputées est un estimateur non biaisé de $P(\underline{X} = \underline{x})$ si le mécanisme donnant lieu à la non-réponse est aléatoire (en anglais, "missing at random data", voir Kalton (1983) ou Rubin (1976)).

En utilisant cette procédure, il est facile de déterminer le nombre espéré de mauvaises imputations pour chaque groupe d'enregistrements ayant le même schéma de non-réponse. Si ce nombre dépasse une valeur fixée d'avance, la procédure n'utilise pas assez d'information et le choix d'un donneur approprié devient quelque peu hasardeux. On peut alors utiliser d'autres variables auxiliaires catégoriques aux enregistrements longitudinaux afin d'améliorer la méthode.

Samuhel et Huggins (1984) proposent également un modèle de régression logistique. Une fois qu'on a obtenu l'ensemble des donneurs possibles pour un enregistrement partiel, on peut dénombrer tous les schémas possibles en assignant à chacun d'eux une valeur unique. On estime ensuite la probabilité d'avoir un schéma particulier à l'aide d'un modèle de régression logistique reliant l'information auxiliaire (variables indépendantes) à la valeur associée à ce schéma. Une fois ces probabilités déterminées, on peut choisir le schéma le plus probable ou choisir un schéma aléatoirement en utilisant la loi de probabilité obtenue.

6.3.2 Méthodes d'imputation pour variables continues

Dans cette section, deux techniques d'imputation pour données continues sont présentées, soit une méthode itérative par la régression ainsi qu'une méthode d'appariement.

6.3.2.1 Méthode d'imputation itérative par régression

Buck (1960) a développé une méthode pour estimer la matrice des covariances de n'importe quelle population multidimensionnelle avec observations manquantes. Beale et Little (1975) utilisent l'estimateur de la matrice des covariances suggéré par Buck pour obtenir de façon itérative les valeurs imputées de même que les estimateurs du vecteur des moyennes et de la matrice de covariance de la population. Cette méthode produit des estimateurs à maximum de vraisemblance si la population est multinormale mais reste quand même valable pour d'autres lois de probabilité.

Supposons qu'on ait n observations et p vagues et dénotons par x_{ij} la valeur d'une variable quelconque X pour la $i^{ième}$ observation à la $j^{ième}$ vague, $i = 1, ..., n$ et $j = 1, ..., p$. Soit μ_j, la moyenne expérimentale portant sur toutes les observations complètes de la $j^{ième}$ vague et σ_{jk}, la covariance expérimentale entre la variable X à la $j^{ième}$ vague et la variable X à la $k^{ième}$ vague portant sur toutes les observations complètes. La méthode itérative de Beale et Little utilise les estimateurs initiaux de Buck μ_j et σ_{jk}, j et k = 1, ..., p, pour calculer

(6.1) $\hat{x}_{ij}$ = x_{ij} si x_{ij} est observé, ou,
une combinaison linéaire des valeurs de la variable X aux vagues observées si x_{ij} est manquant

(6.2) c_{ijk} = la covariance partielle entre x_{ij} et x_{ik} étant donné les valeurs de la variable X aux vagues observées pour la $i^{ième}$ unité si x_{ij} et x_{ik} sont manquants, ou,
0 sinon

Ce qui nous permet d'obtenir

(6.3) $\bar{x}_j = \Sigma_{i=1}^{n} \hat{x}_{ij} / n$, $j = 1, ..., p$

(6.4) $a_{jk} = \Sigma_{i=1}^{n} (\hat{x}_{ij} - \bar{x}_j)(\hat{x}_{ij} - \bar{x}_k) + c_{ij}k$

On pose ensuite $\mu_j = \bar{x}_j$, $\sigma_{jk} = a_{jk}/(n - 1)$ et on répète les étapes (6.1) à (6.4) jusqu'à ce qu'il n'y ait plus de changements significatifs dans les expressions de μ_j et de σ_{jk}, j et k = 1, ..., p. À noter que le terme $c_{ij}k$ de (6.4) est une correction pour le biais.

Si la population n'est pas multinormale, le meilleur estimateur des variables manquantes d'une observation particulière n'est pas nécessairement une combinaison linéaire des variables connues pour cette observation. Ainsi, pour certaines populations, il serait peut-être possible d'obtenir des estimateurs légèrement supérieurs à ceux obtenus par Beale et Little.

Des résultats préliminaires ont indiqué que la méthode s'avérait efficace et performante. Cependant, Beale et Little ne mentionne rien au sujet de la vitesse de convergence de cette méthode itérative.

6.3.2.2 Technique d'appariement

Dénotons par $\underline{y}_i$, le vecteur de dimension 12 représentant les réponses obtenues pour le $i^{ième}$ enregistrement longitudinal partiel. Décomposons ce vecteur en deux sous-vecteurs, $\underline{y}_i(m)$ et $\underline{y}_i(r)$, respectivement les valeurs des mois manquants et des mois observés pour cet enregistrement. Dénotons maintenant par $\underline{x}_j$ les réponses obtenues au $j^{ième}$ enregistrement longitudinal complet. Décomposons de la même façon ce vecteur en deux sous-vecteurs $\underline{x}_j(m)$ et $\underline{x}_j(r)$ correspondant à $\underline{y}_i(m)$ et $\underline{y}_i(r)$. On imputera les valeurs des mois manquants du $i^{ième}$ enregistrement par méthode du plus proche voisin basé sur l'appariement de $\underline{y}_i(r)$ avec les $\underline{x}_j(r)$. Plusieurs distances entre vecteurs peuvent être définies tel que la distance Euclidienne ou la distance de Mahalanobis. On peut choisir, pour l'imputation, le plus proche voisin ou choisir aléatoirement parmi les k plus proches voisins.

6.3.3 Recherche sur la non-réponse de vague dans SIPP

Actuellement, il n'existe pas de fichier longitudinal pour SIPP et la structure de ce dernier n'a pas encore été définie. Seuls les fichiers transversaux ont été traités de façon indépendante considérant chaque vague comme une enquête transversale. Cependant, le premier fichier longitudinal de SIPP contiendra trois interviews, donc douze mois de données, l'unité d'analyse principale étant l'individu. Bailey, Chapman et Kasprzyk (1985) suggèrent de considérer la non-réponse de vague comme la non-réponse totale ou partielle selon les différents schémas de non-réponse possibles. Par exemple, un individu ayant répondu à une seule vague serait considéré comme non-répondant total (corrigé par pondération) alors qu'un individu n'ayant manqué que la deuxième vague serait considéré comme non-répondant partiel (corrigé par imputation).

À partir de la quatrième vague, le questionnaire de SIPP contient une section de "vague manquante" pour les individus n'ayant pas répondu à la vague précédente. Cette section contient une liste de questions rétrospectives portant sur les items importants de la vague précédente. Ceci pourrait permettre dans certains cas de réduire la non-réponse de vague à une non-réponse partielle. Cette section ne s'applique pas à un individu ayant manqué deux vagues ou plus. Une évaluation qualitative de ces questions rétrospectives sera faite avant de les implanter.

Beaucoup de recherches s'effectuent présentement sur le traitement de la non-réponse de vague pour SIPP. Entre autres, Little comparera empiriquement son approche de pondération décrite à la section 2 à des approches de pondération par classe d'ajustement. Kalton, quant à lui, sera amené à comparer le traitement de la non-réponse de vague par pondération et imputation. Pour en savoir plus long à ce sujet, voir l'article de Bailey, Chapman et Kasprzyk.

7. CONCLUSION

Ce travail constitue une synthèse de la littérature existante portant sur le traitement de la non-réponse dans les enquêtes longitudinales. Bien que ce travail porte essentiellement sur la non-réponse de vague, plusieurs méthodes d'imputation pour la non-réponse partielle ont été décrites.

Le sujet de cette étude est en pleine expansion et les méthodes qui y sont analysées sont pour la plupart fort récentes et devraient être testées davantage avant d'être utilisées.

Le dilemme entre imputation et pondération quant au traitement de la non-réponse de vague persiste toujours. En fait, il apparaît souhaitable d'utiliser une combinaison des deux méthodes puisque l'imputation de tous les items des vagues manquantes serait une tâche monumentale et que la pondération pour schémas non emboîtés semble problématique. À cet effet, la suggestion de Little et David paraît très intéressante. Celle-ci consiste à utiliser l'imputation sur les unités n'ayant manqué qu'une seule vague puis d'éliminer, pour les individus restants décrivant un schéma non-emboîté, une partie de l'information obtenue à certaines vagues de façon à obtenir un schéma emboîté.

Finalement, même avec d'autres expérimentations, il est peu probable qu'une méthode soit systématiquement meilleure qu'une autre. La méthode à choisir devra évidemment tenir compte de la nature de l'enquête longitudinale à laquelle on a affaire.

BIBLIOGRAPHIE

Bailey, L., Chapman, D.W. et Kasprzyk, D. (1985). Nonresponse Adjustment Procedures at the Census Bureau: A Review. Proceedings of USBC Research Conference.

Beale, E.M.L. et Little, R.J.A. (1975). Missing Values in Multivariate Analysis. Journal of the Royal Statistical Society. B37:129-146.

Binder, D.A. (1986). Nonresponse and Loglinear Models for Categorical Data. Statistics Canada, Ottawa, Canada.

Buck, S.F. (1960). A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer. Journal of the Royal Statistical Society. B22:302-306.

Cohen, S.B. (1982). An Analysis of Alternative Imputation Strategies for Individuals with Partial Data in the National Medical Care Expenditure Survey. Review of Public Data Use, 10, 153-165.

Cox, B.G. (1980). The Weighted Sequential Hot Deck Imputation Procedure. Proceedings of the American Statistical Association, Survey Research Methods Section, 721-726.

Cox, B.G. (1983). Compensating for Missing Data in the NMCUES. Priorities in Health Statistics 1983, DHHS Publication No. (PHS) 81-1214, National Center for Health Statistics, Hyattsville, Maryland, 186-191.

Cox, B.G. et Cohen, S.B. (1985). Methodological Issues for Health Care Surveys. (Statistics textbooks and monographs; V.61). - Marcel Dekker, Inc., New York et Basel.

Cox, B.G. et Sweetland, S.S. (1982). Imputation of Attrition Related Missing Data for the National Medical Care Utilization and Expenditure Survey. RTI Report No. RTO/1898/06-01F, Contract No. HRA-233-79-2032, Health Care Financing Administration, Baltimore, Maryland and the National Center for Health Statistics, Hyattsville, Maryland.

David, M.R., Little, R.J.A., Samuhel, M.E. et Triest, R.K. (1983). Imputation Models Based on the Propensity to Respond. Draft paper for 1983 American Statistical Association Meetings. Toronto.

Fay, R.E. (1986) Causal Models for Patterns of Nonresponse. Journal of the American Statistical Association, Vol 81, No. 394, Survey Research Methods Section.

Huggins, V.J., Samuhel, M.E. et Weidman, L. (1985). Methodologies for Imputing Longitudinal Survey Items. SIPP Working Paper series.

Kalton, G. (1983).Compensating for Missing Survey Data. Survey Research Center, University of Michigan, Ann Arbor.

Kalton, G. (1985).Handling Nonresponse in Longitudinal Surveys. Paper prepared for USBC Annual Research Conference, Reston, Virginia, March, 1985.

Kalton, G. et Lepkowski, J. (1983). Cross-Wave Item Imputation. In Technical, Conceptual and Administrative Lessons of the Income Survey Development Program (ISDP). M. David ed., 171-198 - Social Science Research Council, Washington, D.C..

Lepkowski, J. (1982). Construction of Longitudinal Data Sets for Waves 1 and 2 of the Income Survey Development Program, 1979 Research Panel. Paper prepared under contract with the Survey Research Center, University of Michigan, 70pp.

Little, R.J.A. (1982).Models for Nonresponse in Sample Surveys. Journal of the American Statistical Association, 77, 237-250.

Little, R.J.A. (1985). Nonresponse Adjustments in Longitudinal Surveys: Models for Categorical Data. Invited paper at the 45th Session of the International Statistical Institute, Amsterdam.

Little, R.J.A. et David, M. (1983). Weighting Adjustments for Nonresponse in Panel Surveys. Working paper. Washington, D.C.: U.S. Bureau of the Census.

Rubin, D.B. (1976). Inference on missing data. Biometrika 63: 581-592.

Samuhel, M.E. et Huggins, V.J. (1984). Longitudinal Item Imputation in a Complex Survey. American Statistical Association. Proceedings of the Survey Research Methods Section. 688-692.